Building data.eurecom.fr

Anne-Elisabeth Gazet

Fall 2011

Introduction

Context

The web as we know it today mainly consists of documents that are linked together. We are now moving to a new era, where raw data is being published on the web as *Linked Data*, and woven into the *Web of Data* or *Semantic Web*.

> "The vision of the Semantic Web is to extend principles of the Web from documents to data. Data should be accessed using the general Web architecture using, e.g., URI-s; data should be related to one another just as documents (or portions of documents) are already. This also means creation of a common framework that allows data to be shared and reused across application, enterprise, and community boundaries, to be processed automatically by tools as well as manually, including revealing possible new relationships among pieces of data."
>
> from the W3C Semantic Web FAQ[1]

In order to achieve these goals, new technologies were developed:

- RDF, for Resource Description Framework, is the main building block of the Semantic Web. It's a simple interchange format, where data is represented as triples of the form (subject, predicate, object) ;
- OWL (Web Ontology Language) and RDFS (RDF Schema) are vocabularies for describing RDF properties and classes ;
- SPARQL (SPARQL Protocol and RDF Query Language) is a protocol and query language for semantic web data sources ;
- SPARQL Update, a companion language for SPARQL, enables modifying a data source.

See [1] if you want to learn more about the Linked Data paradigm.

[1] http://www.w3.org/2001/sw/SW-FAQ

Motivation for the project

There is a trend to increase government transparency by releasing more and more public sector information on the web as linked data. The first initiative of the kind was data.gov.uk in the United Kingdom, which was followed by similar projects around the world.

Universities and schools also generate a lot of data about students, promotions, professors, courses, publications, departments, rooms, schedules, exams, etc. The goal of this project is to take all this data generated by EURECOM, transform it in semantic formats (RDF), interlink it with other data (from other universities and datasets) and publish the whole as linked data in order to develop showcase applications that provide useful services to students and professors.

Contents

Chapter 1

Similar projects

This project was inspired by recent similar initiatives in other universities. I studied them in order to see what kind of data they made available, which approach they took in order to publish their data, and what kind of applications were built thanks to the newly exposed data.

data.open.ac.uk - open linked data from The Open University

The OU's LUCERO (Linking University Content for Education and Research Online) project[1] was the first initiative to expose public information from a university as Linked Open Data. The data.open.ac.uk platform was developed as part of the LUCERO project.

Datasets exposed on this platform include:

Open Research Online: publications from OU researchers ;

OU podcasts: collection of Audio and Video material related to education and research at the Open University ;

Course Descriptions ;

OpenLearn: metadata related to units of teaching and learning material openly available from the OpenLearn website ;

KMi Planet Stories: from the online news system of the Knowledge Media Institute ;

KMi People Profiles.

Vocabularies used to represent the data include:

[1] http://lucero-project.info/lb/

BibO: the Bibliographic Ontology[2] is used to represent information about publications originating from OU researchers ;

W3C Media Ontology: this ontology[3] is used to describe audio and video material in the OU podcasts dataset ;

AIISO and the courseware ontology: the Academic Institution Internal Structure Ontology[4] and the courseware ontology[5] are used to describe courses ;

FOAF: the Friend Of A Friend ontology[6] is used to represent information on the staff members at the Knowledge Media Institute

Applications

- OpenLearn Linked Data[7] makes use of data from data.open.ac.uk to suggest courses, podcasts and other OpenLearn units that relate to an OpenLearn Unit ;
- The OU Expert Search[8] system allows users to find academics at the Open University who are experts in a given domain ;
- OUExperts[9] is a mobile (android) application to find Open Univeristy experts in a given domain, and connect to their social network ;
- Buddy Study[10] suggests potential contacts and Open University courses to follow for students, based on the analysis of the topics in the user's Facebook page.

[2] http://bibliontology.com/specification
[3] http://www.w3.org/TR/mediaont-10/
[4] http://vocab.org/aiiso/
[5] http://courseware.rkbexplorer.com/ontologies/courseware
[6] http://xmlns.com/foaf/spec/
[7] http://fouad.zablith.org/apps/openlearnlinkeddata/
[8] http://kmi-web15.open.ac.uk:8080/ExpertSearchClient/ (accessible inside the OU network only)
[9] http://vimeo.com/19743762
[10] http://www.matthew-rowe.com/BuddyStudy/

data.southampton.ac.uk

data.ox.ac.uk

LODUM - Linked Open Data University of MÃ¼nster

National Research Council (CNR) - data.cnr.it

Conclusion

Chapter 2

Data model and design of our ontology

2.1 Available data

One of the difficulties with this project is that we had to think about the data model before actually getting access to the data. To have a clearer idea of what data was stored at EURECOM, I had a look at the internet and intranet sites of the institute. The information available there is mainly about:

- people (teachers, researchers, doctoral students, staff members),
- research outputs,
- courses.

Assuming these would be the datasets we would access later on, we designed an RDF data model. The aim of this model is to fit the data accurately, as well as to make the data reusable by third parties.

2.2 Choice of external vocabularies

To achieve re-usability, it is paramount to represent data using vocabularies that are used by other projects as well. This way, the data will be easier to consume for existing applications. Furthermore, the fact that a vocabulary was accepted by the community as a de facto standard is a good hint that this vocabulary is well designed, and well fitted to its domain.

Yet, since there are still relatively few universities publishing their data as Linked Data, we could not rely only on popularity to make choices. Besides, each dataset being unique, there are relationships that we wanted to represent, to which we did not find any equivalent in the other projects'

datasets. In this case, we used tools such as Schema-Cache[1], Schemapedia[2] and Sindice[3] to see if there existed any term we could use. Alongside popularity, the other criteria we took into account to select vocabularies are:

- the similitude between the concepts and terms present in the vocabulary, and the concepts needed to represent the data ;
- whether the vocabulary was created as part of a bigger project: this means it is the result of a consensus rather than the vision of a sole individual on the domain.

In the end, we selected the following vocabularies:

- FOAF[4]: the Friend Of A Friend vocabulary defines terms for describing persons and their relations to other people and objects. We use it to describe people at EURECOM.
- DC terms[5]
- Participation[6]: The participation ontology is a simple model for describing the roles that people play within groups. As discussed later, we subclass the class part:Role in the REVE ontology, to describe roles played by people inside EURECOM.
- AIISO[7]: The Academic Institution Internal Structure Ontology provides classes and properties to describe the internal organizational structure of an academic institution. We use the aiiso:Course and aiiso:KnowledgeGrouping classes to define courses and tracks.
- BIBO[8]: The Bibliographic Ontology describe bibliographic things on the semantic Web in RDF. It has been inspired by many existing document description metadata formats. We use it to describe the articles in EURECOM's scientific publications repository.
- LODE[9]: The ontology for Linking Open Descriptions of Events is an ontology for publishing descriptions of historical events as Linked Data, and is designed to be compatible with other event-related vocabularies and ontologies. We use the predicates lode:atTime, lode:atPlace and lode:involvedAgent to describe course sessions.

[1]http://schemacache.com/
[2]http://schemapedia.com/
[3]http://sindice.com/
[4]http://xmlns.com/foaf/spec/
[5]http://dublincore.org/documents/dcmi-terms/
[6]http://vocab.org/participation/schema
[7]http://vocab.org/aiiso/schema
[8]http://bibliontology.com/specification
[9]http://linkedevents.org/ontology/

- OWL-Time[10]
- Rooms[11]

2.3 REVE - the Research and Education Vocabulary for EURECOM

2.4 First data model

2.5 Present state of the data

[10] http://www.w3.org/TR/owl-time/
[11] http://vocab.deri.ie/rooms

Chapter 3

Data conversion

3.1 Accessing the data

3.2 Building RDF triples with Python

3.3 Enriching the data with an external source: GeoNames

3.4 Loading triples in a triplestore using SPARQL Update

Chapter 4

Showcase application : a map of EURECOM publications

Chapter 5

Future work

Bibliography

[1] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.